**Bulk Data Export - Low Level Technical Design**

**5.9** **Bidgely Bulk Data Repatriation (INT29)**

**5.9.1** **INT29(Bulk Data Export – Lifestyle Extract)**

**5.9.1.1**  **Bidgely S3 Bucket to Landing Zone Integration**

Below diagram represents the flow of source data from bidgely S3 Bucket to landing zone for Billing Cycle Itemization. Refer section 2.2 in the Data Lake design document for storage account name per environment.

Graphical user interface, application

Description automatically generated

**Figure 55-Billing Cycle Itemization Bidgely S3 to Landing Zone Ingestion Flow**

Below are the details around each stage in the ingestion flow.

|  |  |
| --- | --- |
| **Flow ID** | **Description** |
| **Flow-1** | One or more gpg encrypted files containing data are copied to “contnspdevdllz/bidgely/int29/inbound”  *Note: container name will change in higher env accordingly*  This flow maps to the step in ADF pipeline where we connect to Bidgely S3 bucket using linked service storing the secrets like access token and secret key in key vault connecting to ADF-01.   * One validation is being done if no files are present for the day(dd) in S3 bucket then fail the pipeline with error message as “No Billing Cycle Itemization record exists for folder “<RES/SMB>/Processed/user/insights/output/userLifestyleProfile/YYYY/MM/DD” |

**5.9.1.2**  **Landing to Raw Zone Integration**

Below diagram represents the flow of source data from landing zone until raw zone for billing cycle itemization. Refer section 2.2 in the Data Lake design document for storage account name per environment.

Graphical user interface, application, Teams

Description automatically generated

**Figure 56- Billing Cycle Itemization Landing Zone to Raw Zone Ingestion Flow**

Below are the details around each stage in the ingestion flow.

|  |  |
| --- | --- |
| **Flow ID** | **Description** |
|  | Below flows map to the steps in the ADF Master Pipeline |
| **Flow -2a** | * The gpg file extension is validated. A deviation from the extension name will be rejected and go to rejects folder in landing zone. * Azure Function will be triggered to decrypt and copy files to “contnspdevdllz/bidgely/int29/wip/<RES/SMB>/” * Files which failed to decrypt will go to rejects folder in landing zone   *Note: container name will change in higher env accordingly*    **Rejection Criteria – Flow 1:**   1. Encrypted gpg file extension to match with extension gpg. A deviation from the extension will be rejected.   Reject file path: “contnspdevdllz /bidgely/int29/rejects” |
| **Flow -2b** | For the business dates when files were present are stored in an array variable and for those dates only copy successfully decrypted gpg files from Inbound to processed folder. |
|  | Encrypted gpg files that fails validation are rejected and moved to  “contnspdevdllz/bidgely/int29/rejects” |
| **Flow -2d** | copy **json** files to   “contnspdevdlrz/ energyinsights/bidgely/billcycleitemization/<RES/SMB>/yyyymmdd\*” |
| **Flow-2c** | Copy the decrypted json files from wip to Backup Storage Container |

Only the decrypted json file containing the data will be retained in the Raw Zone

Landing Zone Reject will contain the gpg files that failed to get processed.

Log file will be created in the Standard Zone to capture the current loads file process status and loaded into the SZ parquet.

**5.9.1.3** **Raw to Standard Zone Integration**

The data load process from Raw Zone to standard zone is a divided based on lifestyle ids. There will be one masteruserLifestyleProfile which will have all the eventid and userid relation which are common to all lifestyle ids.

Text

Description automatically generated

In SZ, folder name is masteruserLifestyleProfile.

For each lifestyle id’s there will different folders in sz which will store the output in parquet only about the particular lifestyle id.

A screenshot of a computer

Description automatically generated with medium confidence

Rejection criteria is only if user\_id, event\_id, last\_updated\_timestamp is null and if user\_id is not in format of 3-part key then it will go raw zone record rejects.

**Figure 60-UserLifestyle Profile Raw Zone to Standard Zone Ingestion Flow**

**Source to Target Mapping**

Source to target mapping for Integration flow 1 and 2 is available in the below location. The mapping file includes the following:

1. Source System Details
2. Data Flow details (ADF Details)
3. Log file name and path
4. Source to standard zone mapping

**5.10.2.4** **ADF Pipeline**

|  |  |  |  |
| --- | --- | --- | --- |
| **Pipeline Name** | **Description** | **Master(M) / Child (C’) indicator** | **Integration** |
| **pl\_bidgely\_int29\_bidgelyS3\_lz** | Pipeline to import files from bidgely S3 bucket to landing zone. | M | 29 |
| **pl\_bidgely\_int29\_lz\_sz** | This pipeline id to decrypt files and move files from landing zone to Raw zone/backup storage and then process files from Raw zone to Standard zone and store the result output file in parquet format. | C | 29 |
| **pl\_durablefunction** | Pipeline in the common directory is used to invoke various functions within function app.   * DecryptDurableFunctionBidgely | C | 29 |
| **pl\_decryptbidgelyfiles** | Pipeline tp decrypt bidgely files using Azure function. | C | 29 |
| **tg\_int29\_<res/smb>** | Triggers for INT29 | C | 29 |

**During 1st time run Load\_Information\_table will have Last\_run\_business\_date as current\_date –1 and pipelineName as “INT29\_<RES/SMB>”**

**So, we must ensure that in s3 bucket there must be folders present for each day until current day.**